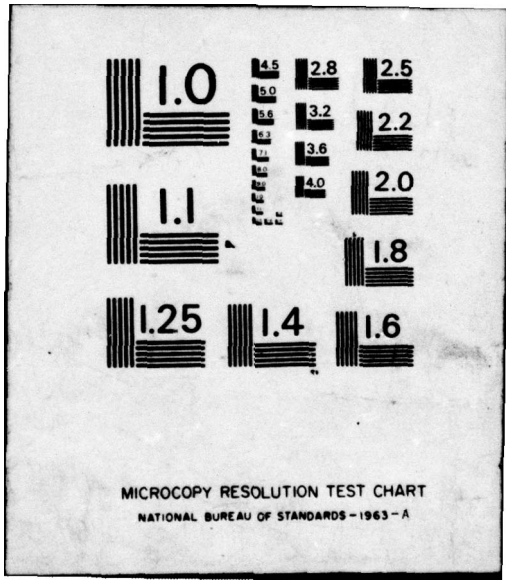
AD-A032 748 SOUTHEASTERN MASSACHUSETTS UNIV NORTH DARTMOUTH DEPT --ETC F/G 12/1
PERFORMANCE BOUNDS OF A CLASS OF SAMPLE-BASED CLASSIFICATION PR--ETC(U)
SEP 76 C H CHEN

UNCLASSIFIED EE-76-6 AFOSR-TR-76-1199 AF-AFOSR-2951-76 NL

| OF |
AD
A032748

END

DATE
FILMED
1-77


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS - 1963 - A

AFOSR - TR - 76 - 1199

AD A032748

TECHNICAL REPORT SERIES IN INFORMATION SCIENCES
(Dr. C. H. Chen, Principal Investigator)

Approved for public release;
distribution unlimited.

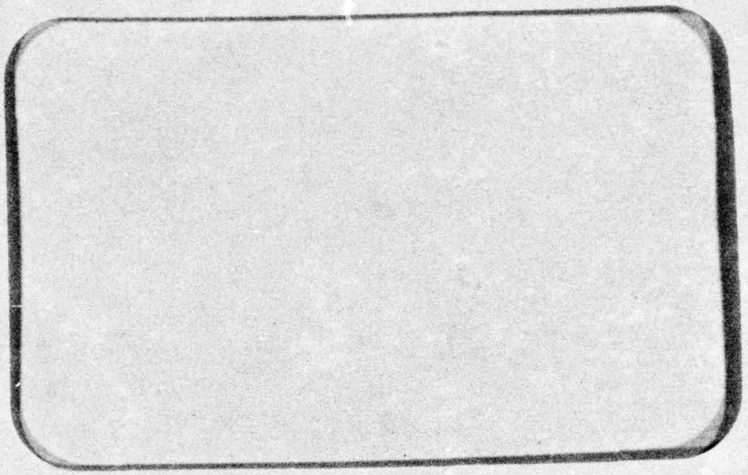
DDC
DEC 2
UG 1932

SOUTHEASTERN MASSACHUSETTS UNIVERSITY

ELECTRICAL ENGINEERING DEPARTMENT

COPY AVAILABLE TO DDC DOES NOT
PERMIT FULLY LEGIBLE PRODUCTION

NORTH DARTMOUTH, MASS. 02747 U.S.A.



AIR FORCE OFFICE OF SCIENTIFIC RESEARCH (AFSC)
NOTICE OF TRANSMITTAL TO DDC
This technical report has been reviewed and is approved for public release IAW AFR 190-12 (7b).
Distribution is unlimited.
A. D. BLOSE
Technical Information Officer

Grant AFOSR 76-2951
TR No. EE-76-6
September 21, 1976

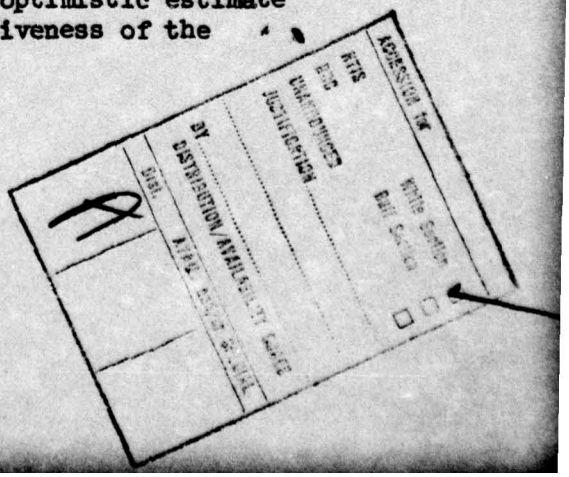
PERFORMANCE BOUNDS OF A CLASS OF SAMPLE-BASED CLASSIFICATION PROCEDURES

by

C. H. Chen
Electrical Engineering Department
Southeastern Massachusetts University
North Dartmouth, Massachusetts 02747

Abstract

Performance bounds of a class of sample-based classification procedures using the k-nearest-neighbor rule (k-NNR) are considered in this paper. By using k-NNR for decision, we show that the lower bounds of the probability of correct decision are very close to that obtained with the Bayes linear discriminant analysis based on the assumption of two multivariate Gaussian densities with different mean vectors but equal covariance matrices. This surprisingly good result suggests that the nonparametric method is very effective at small sample size situation which is of much practical significance. By using the k-NNR for density estimates, an upper bound of the probability of correct decision provides an optimistic estimate of the performance which again indicates the effectiveness of the nonparametric technique.



I. Introduction

This paper is concerned with the performance bounds of a class of sample-based classification procedures using the k-nearest-neighbor rule in decision or density estimate. In particular we show that the lower bounds of the probability of correct decision are very close to that obtained by linear discriminant analysis based on the assumption of two multivariate Gaussian densities with different mean vectors but equal covariance matrices. This surprisingly good result makes the nonparametric methods very attractive at small sample size which is the case in many applications such as pattern recognition, operational research, quality control and related computer science areas.

II. The k-Nearest-Neighbor Classification Rule

Consider two hypotheses H_1 and H_2. Let k be the total number of nearest-neighbors considered. The conditional error of the k-nearest-neighbor rule (k-NNR) for given X is given by Eq. 6-70 of Fukunaga [1],

$$r_k(X) = r^*(X) \sum_{j=0}^{\frac{(k-1)}{2}} \binom{k}{j} r^*(X)^j [1 - r^*(X)]^{k-j} + [1 - r^*(X)] \sum_{j=\frac{(k+1)}{2}}^{k} \binom{k}{j} r^*(X)^j [1 - r^*(X)]^{k-j} \quad (1)$$

where the first and the second terms are the conditional errors for $X \varepsilon H_1$ and $X \varepsilon H_2$ respectively, and $r^*(X)$ is the Bayes conditional error,

$$r^*(X) = \min. [P(H_1/X), P(H_2/X)]$$

It can be shown that Eq. (1) is a concave function of $r^*(X)$. For example, let $k = 3$,

$$r_k(X) = r^*(X)(1 - r^*(X)) \{1 + 4r^*(X) - 4r^*(X)^2\}$$

which is always greater than $r^*(X)$ and is monotonically increasing with $r^*(X)$ for $0 \leq r^*(X) \leq \frac{1}{2}$. $r_k(X)$ is also symmetric with respect to $r^*(X)$. The concave property holds in general for any k. Figure 1 is a typical plot of $r_k(X)$ versus $r^*(X)$ for k = 3.

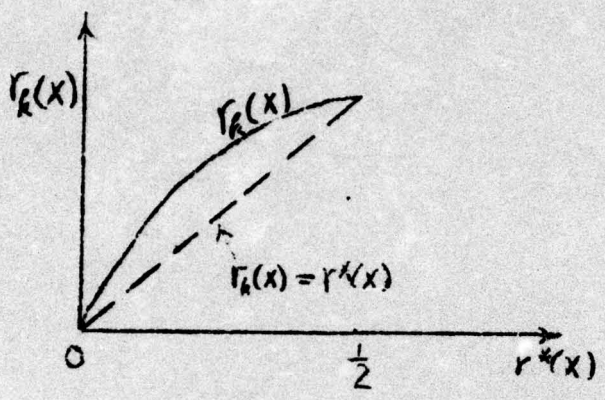


Fig. 1

Taking expectation on both sides of Eq. (1) we obtain the average or unconditional error,

$$r_k = E[r_k(X)] \leq \sum_{j=0}^{\frac{k-1}{2}} \binom{k}{j} [E(r^*(X))]^{j+1} [1 - E(r^*(X))]^{k-j}$$

$$+ \sum_{j=\frac{k+1}{2}}^{k} \binom{k}{j} [E(r^*(X))]^{j} [1 - E(r^*(X))]^{k-j} \qquad (2)$$

Here the Bayes error $E[r^*(X)]$ is unknown, however. By assuming multivariate Gaussian densities for the measurement,

$$p(X/H_1) = n(\mu_1, \Sigma), \quad p(X/H_2) = n(\mu_2, \Sigma)$$

an estimate of the Bayes error is given by (see e.g. [2])

$$G(-\tfrac{D}{2}), \text{ where } G(y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} e^{-t^2/2}\, dt$$

and $D^2 = (\bar{X}_1 - \bar{X}_2)' S^{-1} (\bar{X}_1 - \bar{X}_2)$

is the estimated Mahalanobis distance between the two populations. Here $\bar{X}_1$ and $\bar{X}_2$ are the sample means based on sample sizes n_1 and n_2 respectively and

$$S = \{ \sum_{i=1}^{n_1} (X_i - \bar{X}_1)(X_i - \bar{X}_1)' + \sum_{i=n_1+1}^{n} (X_i - \bar{X}_2)(X_i - \bar{X}_2)' \}/(n - 2)$$

is the sample estimate of the common covariance matrix Σ. $n = n_1 + n_2$ is the total number of training (labelled) samples. Let $\hat{r}_k$ be the sample estimate of the average error,

$$\hat{r}_k = \hat{r}_k^{\,1} + \hat{r}_k^{\,2}$$

where

$$\hat{r}_k^{\,1} = \sum_{j=0}^{\frac{k-1}{2}} \binom{k}{j} G(-\tfrac{D}{2})^{j+1} [1 - G(-\tfrac{D}{2})]^{k-j}$$

$$\hat{r}_k^{\,2} = \sum_{j=\frac{k+1}{2}}^{k} \binom{k}{j} G(-\tfrac{D}{2})^{j} [1 - G(-\tfrac{D}{2})]^{k-j+1}$$

Table 1 is a tabulation of D, $\hat{r}_k^{\,1}$, $\hat{r}_k^{\,2}$, and $\hat{r}_k$. As indicated by Eq. (2), $\hat{r}_k$ is an upper bound of the error probability using k-NNR. $P_c = 1 - \hat{r}_k$ as plotted in Fig. 2 is the lower bound of the probability of correct decision.

Table 1

D	0	1	2	4	6	8
$\hat{r}_k^{\,1}$	0.25	0.255	0.163	0.022	1.3×10^{-3}	3.35×10^{-5}
$\hat{r}_k^{\,2}$	0.25	0.065	0.005	5.4×10^{-8}	2.25×10^{-15}	6.5×10^{-25}
$\hat{r}_k$	0.5	0.32	0.168	0.022	1.3×10^{-3}	3.35×10^{-5}

A proper choice of k depends on the sample size n and the dimension p of the vector measurement. For the Gaussian assumption of densities, the relationship as given by Fukunaga and Hostetler [3] is tabulated in Table 2.

Table 2

p	4	8	16	32
$k_{opt.}$	$0.75n^{1/2}$	$0.94n^{1/3}$	$0.62n^{1/5}$	$0.42n^{1/9}$

It is interesting to note that the above results is consistent with the computer simulation result reported by Goldstein [4], which shows that for $p = 2$, $k_{opt.}$ is proportional to n^{α} with $\alpha \geq 0.5$.

For comparison purpose we shall now consider the performance of linear discriminant function.

III. The Linear Discriminant Analysis

With multivariate Gaussian assumption as described in the previous section, the sample estimate of the error probability $G(-\frac{D}{2})$ is one of several error estimates which have been examined ([2], [5]). $G(-\frac{D}{2})$ in fact provides only a lower bound of the error probability as can be seen below.

McLachlan [2] has given the following expectation of sample error estimate,

$$E[G(-\tfrac{D}{2})] \approx G(-\tfrac{\Delta}{2}) + B_1 + B_2 \tag{4}$$

where Δ^2 is the true Mahalanobis distance when both mean vectors and covariance matrix are known exactly and B_1 and B_2 are given by

$$B_1 = \frac{g(-\frac{\Delta}{2})}{16}\left[\{\Delta - \frac{4(p-1)}{\Delta}\}(\frac{1}{n_1}+\frac{1}{n_2}) + \frac{\Delta}{2}\{\Delta^2 - 4(2p+1)\}\frac{1}{n-2}\right]$$

$$B_2 = \frac{g(-\frac{\Delta}{2})}{1024}\Big([\Delta\{\Delta^2 - 4(2p+1)\} + \frac{16(p-1)}{\Delta}\{(p-1) + \frac{4(p-3)}{\Delta^2}\}](\frac{1}{n_1}+\frac{1}{n_2})^2$$

$$+ \{\Delta^5 - 4(3p+7)\Delta^3 + 16(2p^2+8p+5)\Delta - \frac{64(p-1)(2p+1)}{\Delta}\}(\frac{1}{n_1}+\frac{1}{n_2})(\frac{1}{n-2})$$

$$+ \frac{\Delta}{12}\{3\Delta^6 - 4(12p+35)\Delta^4 + 16(12p^2+72p+71)\Delta^2$$

$$- 192(12p^2+12p+1)\}\frac{1}{(n-2)^2}\Big)$$

where $g(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2}$. Both B_1 and B_2 approach 0 as $n_1, n_2 \to \infty$. B_1 obviously is not good for small Δ. A sample calculation for $n_1 = n_2 = 5$ and $p = 2$ gives $B_1 = -0.030$, $B_2 = -0.00118$ at $\Delta = 2$, and $B_1 = -0.000675$ and $B_2 = 0$ at $\Delta = 4$. Since B_1 and B_2 are negative or otherwise negligibly small, $E[G(-\frac{D}{2})]$ is thus smaller than its true value. $1 - G(-\frac{D}{2})$ is also plotted in Figure 2. The close proximity of the two curves clearly indicates that the k-NNR can provide a performance very close to the Bayes decision rule under Gaussian assumption.

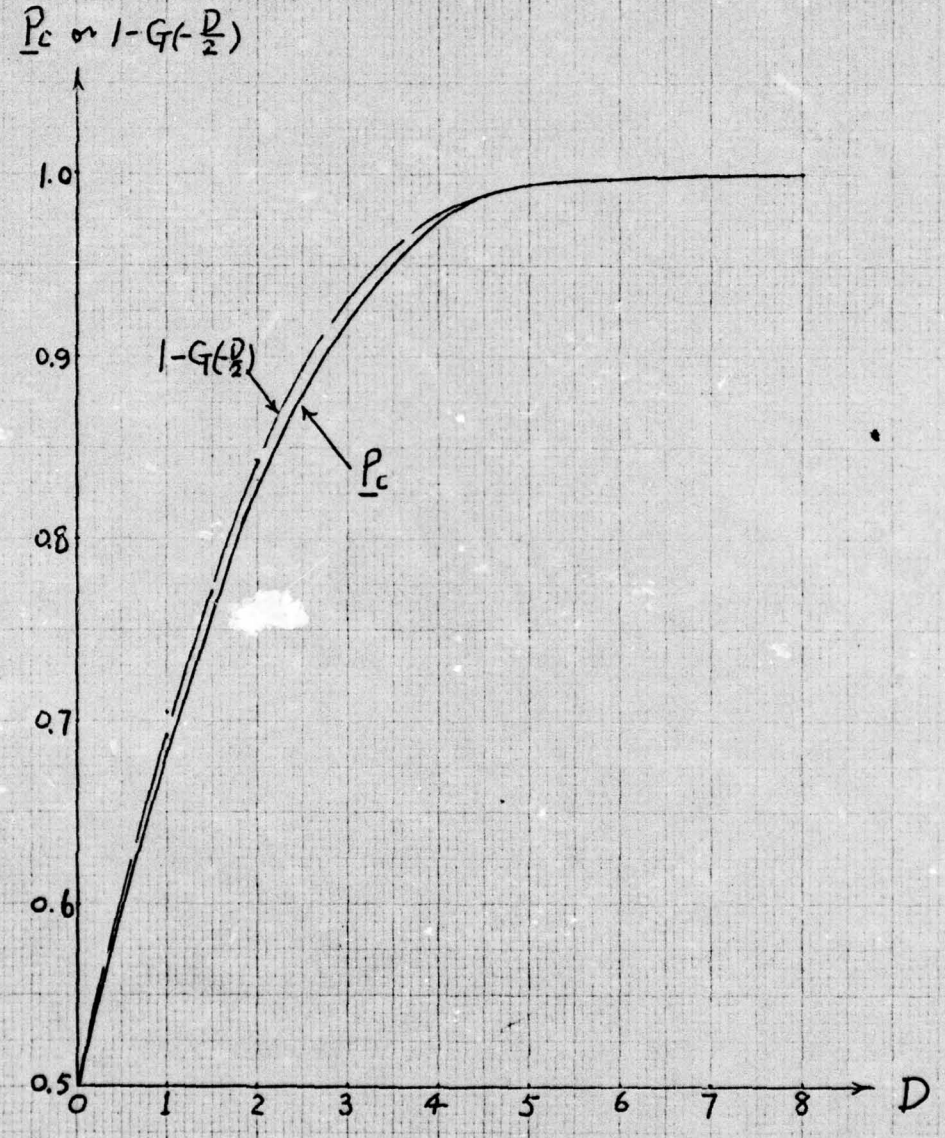


Fig. 2

20 Squares to the Inch

IV. Classification Based on the k-Nearest-Neighbor Density Estimates

Let k_1 and k_2 be the numbers of nearest neighbors from populations H_1 and H_2 respectively; $k = k_1 + k_2$. Let d_k be the distance between X and its kth nearest-neighbor, nearness being measured by any convenient metric $d(X, Y)$. Let $S_i(X)$ be the region about X containing its k-NN,

$$S_i(X) = \{Y: \ d_i(X, Y) \leq d_k\}, \quad i = 1,2$$

where the index refers to the hypothesis considered. Also let $v_i(X)$ be the volume of this region,

$$v_i(X) = \int_{S_i(X)} dY$$

The k-NN estimate of the probability density is

$$\hat{p}_i(X) = \frac{k_i - 1}{n_i v_i(X)}$$

We can now define new test statistics as

$$t_1 = \frac{k_1 - 1}{nn_1} \sum_{i=1}^{n} \frac{1}{v_1(X_i)}; \qquad t_2 = \frac{k_2 - 1}{nn_2} \sum_{i=1}^{n} \frac{1}{v_2(X_i)} \tag{5}$$

and the decision rule is to accept hypothesis H_1 if $t_1 \geq t_2$, otherwise accept hypothesis H_2. Also define the coverage $u_i(X)$ as

$$u_i(X) = \int_{S_i(S)} p_i(Y)dY; \quad p_i(Y) = p(Y/H_i)$$

It has been shown [3] that

$$\frac{1}{v_i(X)} \simeq \frac{p_i(X)}{u_i(X)} + \frac{C_i(X)}{p_i(X)^{2/p}\, u_i^{1-2/p}(X)}$$

where

$$C_i(X) = \frac{1}{2(p + 2)\pi} \Gamma^{2/p} \left(\frac{p + 2}{2}\right) t\gamma\left\{\left[\frac{A_i}{|A_i|^{1/p}}\right]^{-1} \left[\frac{\partial^2 p_i(X)}{\partial X^2}\right]\right\}$$

where A_i is the transformation matrix for

$$d_i^2(X, Y) = (Y - X)'A_i(Y - X).$$

Now the mean and variance of t_i can be determined as:

$$\bar{t}_i = E[t_i] \simeq \frac{k_i - 1}{n_i} \left\{ \frac{n}{k-1} p_i(X) + \left(\frac{n}{k-1}\right)^{1-2/p} C_i(X) p_i^{-2/p}(X) \right\}$$

$$\sigma_{t_i}^2 = E[t_i - \bar{t}_i]^2 \simeq \left(\frac{k_i - 1}{n_i}\right)^2 \frac{n(1+n-k)}{(k-1)^2(k-2)} p_i^2(X) \qquad (6)$$

It is noted that the mean and variance are functions of X because the expectation was taken with respect to u instead of X. Assuming that t_i is univariate Gaussian distributed, the probability of correct descision, for given X, is given by

$$\text{Prob.}\ (t_1 > t_2 | H_1, X)$$

$$= \int_{-\infty}^{\infty} \left[\int_{t_2}^{\infty} \frac{1}{\sqrt{2\pi}\,\sigma_{t_1}} e^{-\frac{(y-\bar{t}_1)^2}{2\sigma_{t_1}^2}} dy \right] \frac{1}{\sqrt{2\pi}\,\sigma_{t_2}} e^{-\frac{(t_2-\bar{t}_2)^2}{2\sigma_{t_2}^2}} dt_2$$

$$= \int_{-\infty}^{\infty} \left[\int_{\frac{\bar{t}_2 - \bar{t}_1 + \sigma_{t_2}\omega}{\sigma_{t_1}}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \right] \frac{1}{\sqrt{2\pi}} e^{-\omega^2/2} d\omega \qquad (7)$$

Although the average probability of correct decision may be obtained by taking the expectation of Eq. (7) with respect to X, the resulting expression is difficult to evaluate. If, however, we assume both $p_1(X)$ and $p_2(X)$ are multivariate Gaussian with sample estimates of means and common covariance matrix as described in previous section, then an upper bound of the probability of correct decision may be obtained by evaluating Eq. (7) at $X = \bar{X}_1$. By neglecting the second term for $\bar{t}_i$ in Eq. (6), a sample plot of the probability bound of correct decision versus the estimated Mahalanobis distance is shown in Fig. 3.

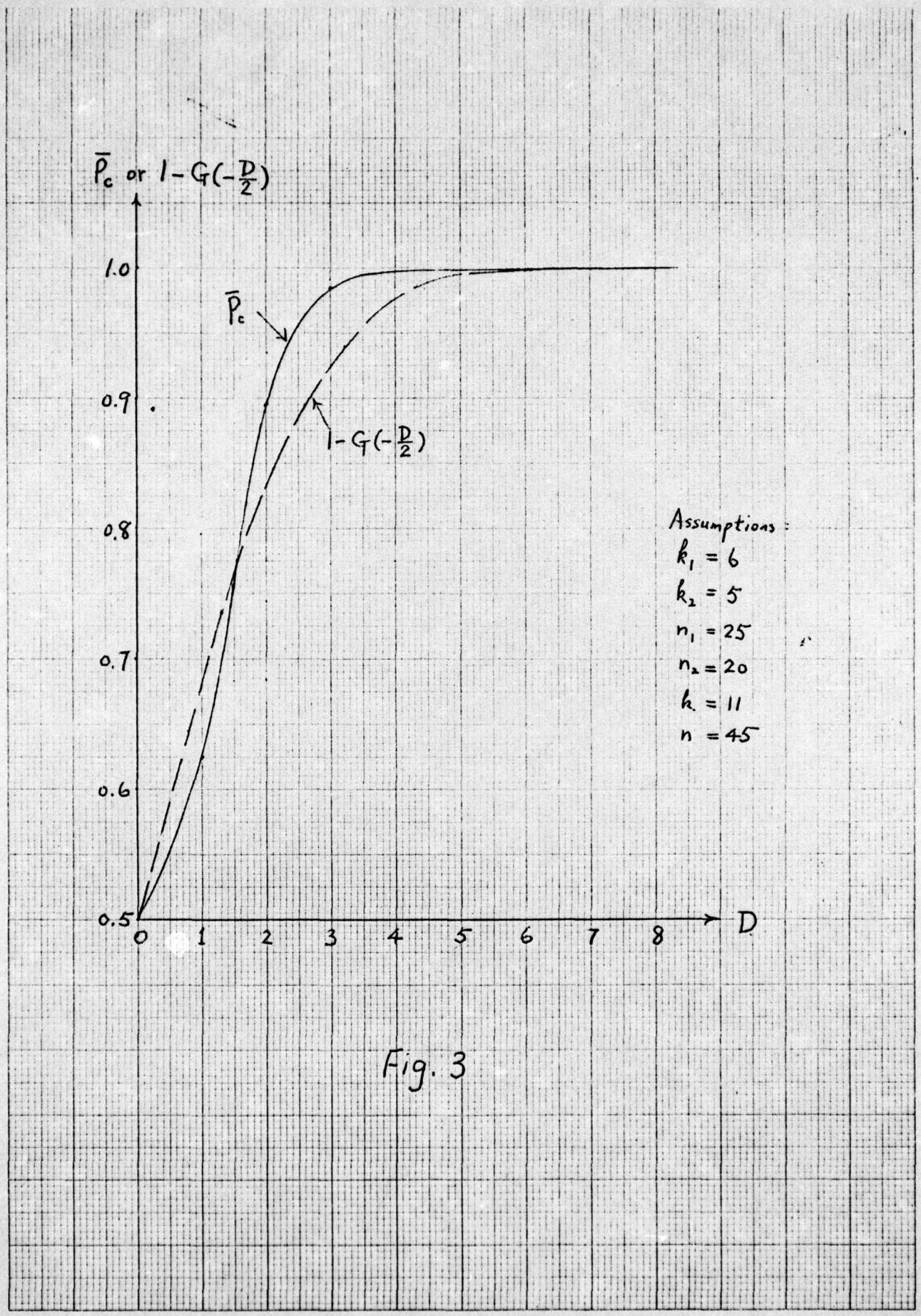

$\bar{P}_c$ or $1-G(-\frac{D}{2})$
1.0
0.9
0.8
0.7
0.6
0.5
$\bar{P}_c$
$1-G(-\frac{D}{2})$
Assumptions:
$k_1 = 6$
$k_2 = 5$
$n_1 = 25$
$n_2 = 20$
$k = 11$
$n = 45$
0
1
2
3
4
5
6
7
8
D

Fig. 3

References

1. K. Fukunaga, "Introduction to Statistical Pattern Recognition", Academic Press, Inc., 1972.

2. G.J. McLachlan, "Esimate of Errors of Misclassification on the Criterion of Asymptotic Mean Square Error", Technometrics, Vol. 16, No. 2, pp. 255-260, May 1974.

3. K. Fukunaga and L.D. Hostetler, "Optimization of k-Nearest-Neighbor Density Estimators", IEEE Trans. on Information Theory, Vol. IT-19, No. 3, pp. 320-326, May 1973.

4. M. Goldstein, "Comparison of Some Density Estimate Classification Procedures", Journal of the American Statistical Association, Vol. 70, No. 351, pp. 666-669, September 1975.

5. P.A. Lachenbruch and M.R. Mickey, "Estimation of Error Rates in Discriminant Analysis", Technometrics, Vol. 10, No. 1, pp. 1-11, February 1968.

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

(18) (19)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: AFOSR - TR - 76 - 1199	2. GOVT ACCESSION NO.
	3. RECIPIENT'S CATALOG NUMBER: (9) Technical Rept.
4. TITLE (and Subtitle): (6) PERFORMANCE BOUNDS OF A CLASS OF SAMPLE-BASED CLASSIFICATION PROCEDURES.	5. TYPE OF REPORT & PERIOD COVERED: Interim
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): (10) C. H. Chen (11) 21 Sep 76	8. CONTRACT OR GRANT NUMBER(s): (14) AFOSR 76-2951 EE-76-6
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Southeastern Massachusetts University, Electrical Engineering Department, North Dartmouth, MA 02747 (12) 10p.	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 61102F 2304/A2
11. CONTROLLING OFFICE NAME AND ADDRESS: Air Force Office of Scientific Research/NM, Bolling AFB, Washington, DC 20332	12. REPORT DATE: September 21, 1976
	13. NUMBER OF PAGES: 9
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office): (15) AF-AFOSR-2951-76	15. SECURITY CLASS. (of this report): UNCLASSIFIED
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report): (16) 2304 (17) A2

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Performance bounds of a class of sample-based classification procedures using the k-nearest-neighbor rule (k-NNR) are considered in this paper. By using k-NNR for decision, we show that the lower bounds of the probability of correct decision are very close to that obtained with the Bayes linear discriminant analysis based on the assumption of two multivariate Gaussian densities with different mean vectors but equal covariance matrices. This surprisingly good result suggests that the nonparametric method is very effective at small sample size situation which is of much practical signficance. By using the k-NNR →next page

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

407932

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20 Abstract

for density estimates, an upper bound of the probability of correct decision provides an optimistic estimate of the performance which again indicates the effectiveness of the nonparametric technique.